d his

(FILE 'HOME' ENTERED AT 11:43:00 ON 03 DEC 2003)

FILE 'CAPLUS' ENTERED AT 11:47:26 ON 03 DEC 2003
L1 116 S ((PEPTIDE OR PROTEIN) (2W) DATABASE) AND FUNCTION## AND RESID

=> d bib,abs 12,29,30,40,41,42,46,53,57,63,69,70,80,82,87,101

L1 ANSWER 12 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2003:181872 CAPLUS
TI Pattern recognition tools for protein sequence analysis
AU Lybrand, Terry P.; Li, Zhijun
CS Department of Chemistry and Center for Structural Biology, Vanderbilt University, Nashville, TN, 37235, USA
SO Abstracts of Papers, 225th ACS National Meeting, New Orleans, LA, United States, March 23-27, 2003 (2003), COMP-186 Publisher: American Chemical Society, Washington, D. C.
CODEN: 69DSA4
DT Conference; Meeting Abstract
LA English
AB The rapid growth in **protein** sequence **databases** provides a wealth of information about sequence-**function** relationships. For example, multiple sequence alignments for functionally related protein families often exhibit sequence patterns, or motifs, that may reveal interesting information about the location and nature of ligand binding sites or other important features. These motifs may be complex, consisting of multiple **residues** well sepd. in sequence space. Hence, effective pattern recognition tools are needed to facilitate sequence motif identification. We report here a simple, but robust, set of procedures and tools to aid in identification of complex sequence motifs in large multiple sequence alignment data sets. Several test cases are presented, where exptl. data are available to confirm the significance of **residues** identified in interesting motifs. We will also discuss extension of the method to other pattern recognition tasks, such as anal. of small mol. data sets for interesting **functional** group substitution patterns.

L1 ANSWER 29 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2002:46481 CAPLUS
DN 137:181846
TI Development of a database for protein cavities and its usage for similarity searches in binding sites
AU Schmitt, Stefan; Hendlich, Manfred; Klebe, Gerhard
CS Department of Pharmaceutical Chemistry, Philipps-Universitat Marburg, Marburg, D-35032, Germany
SO Rational Approaches to Drug Design, Proceedings of the European Symposium on Quantitative Structure-Activity Relationships, 13th, Duesseldorf, Germany, Aug. 27-Sept. 1, 2000 (2001), Meeting Date 2000, 135-141. Editor(s): Hoeltje, Hans-Dieter; Sippl, Wolfgang. Publisher: Prous Science, Barcelona, Spain.
CODEN: 69CEP6; ISBN: 84-8124-176-8
DT Conference
LA English
AB A new database was developed for a similarity anal. of protein cavities using a clique detection method followed by a scoring of the resulting cliques. The cavity description was reduced to a set of representative pseudo centers that comprised the physicochem. properties of the margin **residues** flanking the cavity. This method was capable to retrieve from a sample of 5.500 cavities, sets of functionally related proteins, including reliable detection of cavities accommodating similar ligands or ligand portions. It also detected on the most prominent rankings the active site of subtilisin to be similar with thrombin. These results demonstrated the strength of this method and indicated its predictive

power to correctly detect **functional** relationships among proteins of non-significant sequence and folding homol.
RE.CNT 17 THERE ARE 17 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 30 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2002:46478 CAPLUS
DN 136:179796
TI Data mining protein interactions: New examples of the serine protease inhibitor canonical loop conformation found in extracellular proteins
AU Jackson, R. M.; Russell, R. B.
CS Department of Biochemistry and Molecular Biology, University College, London, WC1E 6BT, UK
SO Rational Approaches to Drug Design, Proceedings of the European Symposium on Quantitative Structure-Activity Relationships, 13th, Duesseldorf, Germany, Aug. 27-Sept. 1, 2000 (2001), Meeting Date 2000, 105-114. Editor(s): Hoeltje, Hans-Dieter; Sippl, Wolfgang. Publisher: Prous Science, Barcelona, Spain.
CODEN: 69CEP6; ISBN: 84-8124-176-8
DT Conference
LA English
AB Methods for the prediction of protein **function** from structure are of growing importance in the age of structural genomics. Here we focus on the problem of identifying sites of potential serine protease inhibitor interactions on the surface of proteins of known structure. Given that there is no sequence conservation within canonical loops from different inhibitor families we first compare representative loops to all fragments of equal length among proteins of known structure by calcg. main-chain RMS deviation. Fragments with RMS deviation below a certain threshold (hits) are removed if **residues** have solvent accessibilities appreciably lower than those obsd. in the search structure. These remaining hits are further filtered to remove those occurring largely within secondary structure elements. Likely **functional** significance is restricted further by considering only extracellular protein domains. By comparing different canonical loop structures to the **protein** structure **database** we show that the method was able to detect previously known inhibitors. In addn., we discuss potentially new canonical loop structures found in secreted hydrolases, toxins, viral proteins, cytokines and other proteins. We discuss the possible **functional** significance of several of the examples found.
RE.CNT 18 THERE ARE 18 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 40 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2001:20971 CAPLUS
DN 134:291667
TI Structural/**functional** assignment of unknown bacteriophage T4 **proteins** by iterative **database** searches
AU Kawabata, T.; Arisaka, F.; Nishikawa, K.
CS Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan
SO Gene (2000), 259(1-2), 223-233
CODEN: GENED6; ISSN: 0378-1119
PB Elsevier Science B.V.
DT Journal
LA English
AB Among the total of 274 orfs within bacteriophage T4, only half have been reasonably well characterized, and the **functions** of the rest have remained obscure. In order to predict the mol. **functions** of the orfs, a position-specific iterated (PSI)-BLAST search of bacteriophage T4 against the sequence database of known 3D structures was carried out. PSI-BLAST is one of the most powerful iterative sequence search methods using multiple sequence alignment, with the ability to

detect many more proteins with distant homol. than std. pairwise methods. The 3D structures of proteins are considered to be better preserved than the sequences, and the detected distantly homologous proteins are likely to possess highly similar 3D structures. Thirteen orfs of phage T4, whose homologs were not detected by std. pairwise methods, were found to have significantly homologous counterparts by this method. The plausibility of the results was confirmed by checking whether important **residues** at substrate/ligand-binding sites were conserved. Among them, two orfs, vs.1 and e.1, which are similar to Escherichia coli lytic enzyme and MutT protein, resp., had not been studied previously. Also, gp rIIA, a rapid lysis protein, whose gene structure had been intensively studied during the development of mol. biol. in the 1950s and yet whose mol. **function** remains unknown, has an N-terminal domain that is significantly similar to the N-terminal region of the heat shock protein Hsp90.

RE.CNT 46 THERE ARE 46 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 41 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2000:891656 CAPLUS
DN 134:1699
TI Introduction to the thermodynamic database for proteins and mutants, ProTherm
AU Uedaira, Hatsuho; Gromiha, Michael M.; An, Jianghong; Sarai, Akinori
CS Tsukuba Inst., The Inst. Phys. Chem. Res. (RIKEN), Japan
SO Netsu Sokutei (2000), 27(5), 250-256
CODEN: NESOD2; ISSN: 0386-2615
PB Nippon Netsu Sokutei Gakkai
DT Journal; General Review
LA Japanese
AB A review with 12 refs. Thermodn. properties of proteins are essential for understanding the mechanism of their folding and stability. We introduce and explain the database, "ProTherm: Thermodn. Database for Proteins and Mutants", which was developed in our lab. and can be searched through WWW on various conditions with different options for output. The database contains more than 7000 numerical data for several important thermodn. properties, structural information of proteins and mutants, exptl. methods and conditions, **functional** and literature information. The database is cross-linked with PubMed, PDB, PMD, EC, and 3DinSight which is developed in our lab. The mutation sites and surrounding **residues** are automatically mapped on structures and can be directly viewed through 3DinSight. ProTherm can be accessed through the World Wide Web at: http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html.

L1 ANSWER 42 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2000:880061 CAPLUS
DN 134:277558
TI Protein **functional**-group 3D motif and its applications
AU Ye, Yuzhen; Xie, Tao; Ding, Dafu
CS Shanghai Institute of Biochemistry, Chinese Academy of Sciences, Shanghai, 200031, Peop. Rep. China
SO Chinese Science Bulletin (2000), 45(22), 2044-2052
CODEN: CSBUEF; ISSN: 1001-6538
PB Science in China Press
DT Journal
LA English
AB Representing and recognizing protein active sites sequence motif (1D motif) and structural motif (3D motif) is an important topic for predicting and designing protein **function**. Prevalent methods for extg. and searching 3D motif always consider **residue** as the minimal unit, which have limited sensitivity. Here we present a new spatial representation of protein active sites, called "**functional** -group 3D motif", based on the fact that the **functional** groups inside a **residue** contribute mostly to its **function**.

Relevant algorithm and computer program are developed, which could be widely used in the **function** prediction and the study of structural-**function** relationship of proteins. As a test, we defined a **functional**-group 3D motif of the catalytic triad and oxyanion hole with the structure of porcine trypsin (PDB code: 1mct) as the template. With our motif-searching program, we successfully found similar sub-structures in trypsins, subtilisins and .alpha./.beta. hydrolases, which show distinct folds but share similar catalytic mechanism. Moreover, this motif can be used to elucidate the structural basis of other proteins with variant catalytic triads by comparing it to those proteins. Finally, we scanned this motif against a non-redundant **protein** structure **database** to find its matches, and the results demonstrated the potential application of **functional** group 3D motif in **function** prediction. Above all, compared with the other 3D-motif representations on **residues**, the **functional** group 3D motif achieves better representation of protein active region, which is more sensitive for protein **function** prediction.

RE.CNT 22 THERE ARE 22 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 46 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2000:592163 CAPLUS
DN 133:278306
TI Identifying sequence-structure pairs undetected by sequence alignments
AU Miyazawa, Sanzo; Jernigan, Robert L.
CS Faculty of Technology, Gunma University, Gunma, 376, Japan
SO Protein Engineering (2000), 13(7), 459-475
CODEN: PRENE9; ISSN: 0269-2139
PB Oxford University Press
DT Journal
LA English
AB We examine how effectively simple potential **functions** previously developed can identify compatibilities between sequences and structures of **proteins** for **database** searches. The potential **function** consists of pairwise contact energies, repulsive packing potentials of **residues** for overly dense arrangement and short-range potentials for secondary structures, all of which were estd. from statistical preferences obsd. in known protein structures. Each potential energy term was modified to represent compatibilities between sequences and structures for globular proteins. Pairwise contact interactions in a sequence-structure alignment are evaluated in a mean field approxn. on the basis of probabilities of site pairs to be aligned. Gap penalities are assumed to be proportional to the no. of contacts at each **residue** position, and as a result gaps will be more frequently placed on protein surfaces than in cores. In addn. to min. energy alignments, we use probability alignments made by successively aligning site pairs in order by pairwise alignment probabilities. The results show that the present energy **function** and alignment method can detect well both folds compatible with a given sequence and, inversely, sequences compatible with a given fold, and yield mostly similar alignments for these two types of sequence and structure pairs. Probability alignments consisting of most reliable site pairs only can yield extremely small root mean square deviations, and including less reliable pairs increases the deviations. Also, it is obsd. that secondary structure potentials are usefully complementary to yield improved alignments with this method. Remarkably, by this method some individual sequence-structure pairs are detected having only 5-20% sequence identity.

RE.CNT 51 THERE ARE 51 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 53 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 2000:191947 CAPLUS
DN 132:319025

TI TOP: a new method for protein structure comparisons and similarity searches
AU Lu, Guoguang
CS Division of Molecular Structural Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, 17177, Swed.
SO Journal of Applied Crystallography (2000), 33(1), 176-183
CODEN: JACGAR; ISSN: 0021-8898
PB Munksgaard International Publishers Ltd.
DT Journal
LA English
AB In order to facilitate the three-dimensional structure comparison of proteins, software for making comparisons and searching for similarities to **protein** structures in **databases** has been developed. The program identifies the **residues** that share similar positions of both main-chain and side-chain atoms between two proteins. The unique **functions** of the software also include database processing via Internet- and Web-based servers for different types of users. The developed method and its friendly user interface copes with many of the problems that frequently occur in protein structure comparisons, such as detecting structurally equiv. **residues**, misalignment caused by coincident match of C.alpha. atoms, circular sequence permutations, tedious repetition of access, maintenance of the most recent database, and inconvenience of user interface. The program is also designed to cooperate with other tools in structural bioinformatics, such as the 3DB Browser software, for convenient mol. modeling and protein structure anal. A similarity ranking score of "structure diversity" is proposed in order to est. the evolutionary distance between proteins based on the comparisons of their three-dimensional structures. The **function** of the program has been utilized as a part of an automated program for multiple protein structure alignment. In this paper, the algorithm of the program and results of systematic tests are presented and discussed.
RE.CNT 42 THERE ARE 42 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 57 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1999:717031 CAPLUS
DN 132:49566
TI Filtered neighbors threading
AU Bienkowska, Jadwiga R.; Rogers, Robert G., Jr.; Smith, Temple F.
CS BioMolecular Engineering Research Center, College of Engineering, Boston University, Boston, MA, 02215, USA
SO Proteins: Structure, Function, and Genetics (1999), 37(3), 346-359
CODEN: PSFGEY; ISSN: 0887-3585
PB Wiley-Liss, Inc.
DT Journal
LA English
AB A knowledge-based threading scoring **function** that exploits the information about protein structure contained in **residue** packing/neighbor preferences is presented. The proposed algorithm eliminates the stereochem. improbable phys. contacts for each possible sequence-to-structure alignment. This algorithm was used to filter the score of the sequence-to-structure alignment. The set of neighbor pairs contributing to the alignment score varies during threading. Whether or not a neighbor pair contributes to the score depends on the threaded amino acids. A detailed structure description that encodes amino acid side-chain rotamer and phys. contact preferences but does not imprint the fold model with the native sequence or native phys. contacts, is used. This description is discretized to collect accurate statistics for the scoring **function** generation. The original detailed description for the neighbor filtering is used. On av., the filtered neighbors threading (FNT) method predicts the sequence-to-structure alignment twice as accurately as does the std. unfiltered neighbors threading. For the set of threadings tested by the PHDthreader method, the FNT gives

predictions with a sequence-to-structure alignment accuracy of 46.9%, which amts. to a 754% improvement in alignment sensitivity compared with PHDthreader predictions. Results show that redn. of noise from the obsd. neighbor pair preferences by filtering leads to noticeable improvements in the predicted sequence-to-structure alignments.
RE.CNT 46 THERE ARE 46 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 63 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1999:298232 CAPLUS
DN 131:213749
TI Protein sequences and genome databases
AU Mewes, Hans-Werner; Maierl, Andreas; Frishman, Dimitrij
CS MPI for Biochemistry, Martinsried, D-82152, Germany
SO Microcharacterization of Proteins (2nd Edition) (1999), 301-317. Editor(s): Kellner, Roland; Lottspeich, Friedrich; Meyer, Helmut E. Publisher: Wiley-VCH Verlag GmbH, Weinheim, Germany.
CODEN: 67PJAN
DT Conference
LA English
AB The advent of the Internet has revolutionized the method of access to protein sequences and genome databases. Data in the biol. sequence database are generally characterized as facts and assocd., interpretive information. Mol. sequence databases provide an extraordinary resource to gain insight into biol. **function**. They serve as repositories for exptl. results as well as ref. compendia, summarizing the current state of biol. knowledge. In principle, they are structured in the form of sep. entries, representing unbranched chains of nucleic acid or amino acid **residues**.
RE.CNT 70 THERE ARE 70 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 69 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1998:318707 CAPLUS
DN 129:108676
TI Application of a novel and fast information-theoretic method to the discovery of higher-order correlations in **protein databases**
AU Steeg, Evan W.; Pham, Hai
CS Molecular Mining Corporation, PARTEQ Innovations, Queen's University, Kingston, ON, K7L 3N6, Can.
SO Pacific Symposium on Biocomputing '98, Maui, Hawaii, Jan. 4-9, 1998 (1998), 573-584. Editor(s): Altman, Russ B. Publisher: World Scientific, Singapore, Singapore.
CODEN: 66CDAZ
DT Conference
LA English
AB We present a fast, discrete data-mining approach to the problem of finding .kappa.-tuples of correlated amino acid **residues** in protein sequence data. When sets of sequence-distant sites display high mutual information, they may bespeak important structural or **functional** features. Our novel methodol. overcomes the limitations of previous methods which examd. only single-**residue** features or pairwise interactions.
RE.CNT 23 THERE ARE 23 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 70 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1998:288379 CAPLUS
DN 129:38341
TI Influence of **protein** structure **databases** on the predictive power of statistical pair potentials
AU Furuichi, Emiko; Koehl, Patrice
CS CNRS, Illkirch Graffenstaden, 67400, Fr.

SO Proteins: Structure, Function, and Genetics (1998), 31(2), 139-149
CODEN: PSFGEY; ISSN: 0887-3585
PB Wiley-Liss, Inc.
DT Journal
LA English
AB A long standing goal in protein structure studies is the development of reliable energy **functions** that can be used both to verify protein models derived from exptl. constraints as well as for theor. protein folding and inverse folding computer expts. In that respect, knowledge-based statistical pair potentials have attracted considerable interests recently mainly because they include the essential features of protein structures as well as solvent effects at a low computing cost. However, the basis on which statistical potentials are derived have been questioned. In this paper, we investigate statistical pair potentials derived from protein three-dimensional structures, addressing in particular questions related to the form of these potentials, as well as to the content of the database from which they are derived. We have shown that statistical pair potentials depend on the size of the proteins included in the database, and that this dependence can be reduced by considering only pairs of **residue** close in space (i.e., with a cutoff of 8 .ANG.). We have shown also that statistical potentials carry a memory of the quality of the database in terms of the amt. and diversity of secondary structure it contains. We find, for example, that potentials derived from a database contg. .alpha.-proteins will only perform best on .alpha.-proteins in fold recognition computer expts. We believe that this is an overall weakness of these potentials, which must be kept in mind when constructing a database.
RE.CNT 43 THERE ARE 43 CITED REFERENCES AVAILABLE FOR THIS RECORD
ALL CITATIONS AVAILABLE IN THE RE FORMAT

L1 ANSWER 80 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1996:471616 CAPLUS
DN 125:159638
TI Pairwise and searchwise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames
AU Birney, Ewan; Thompson, Julie D.; Gibson, Toby J.
CS European Molecular Biology Laboratory, Heidelberg, D-69012, Germany
SO Nucleic Acids Research (1996), 24(14), 2730-2739
CODEN: NARHAD; ISSN: 0305-1048
PB Oxford University Press
DT Journal
LA English
AB DNA translation frames can be disrupted for several reasons, including: (i) errors in sequence detn.; (ii) RNA processing, such as intron removal and guide RNA editing; (iii) less commonly, polymerase frameshifting during transcription or ribosomal frameshifting during translation. Frameshifts frequently confound computational activities involving homologous sequences, such as database searches and inferences on structure, **function** or phylogeny made from multiple alignments. A dynamic alignment algorithm is reported here which compares a protein profile (a **residue** scoring matrix for one or more aligned sequences) against the three translation frames of a DNA stand, allowing frameshifting. The algorithm has been incorporated into a new package, WiseTools, for comparison of biol. sequences. A protein profile can be compared against either a DNA sequence or a protein sequence. The program PairWise may be used interactively for alignment of any two sequence inputs. SearchWise can perform combinations of searches through DNA or **protein databases** by a protein profile or DNA sequence. Routine application of the programs has revealed a set of database entries with frameshifts caused by errors in sequence detn.

L1 ANSWER 82 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1996:210800 CAPLUS
DN 124:282374

TI Search for ancient patterns in protein sequences
AU Thode, Guillermo; Garcia-Ranea, Juan Antonio; Jimenez, Juan
CS Facultad Ciencias, Univ. Malaga, Malaga, 29071, Spain
SO Journal of Molecular Evolution (1996), 42(2), 224-33
CODEN: JMEVAU; ISSN: 0022-2844
PB Springer
DT Journal
LA English
AB Proteins of related **functions** are often similar in sequence, reflecting a common phylogenetic origin. Proteins with no known homol. are probably diversified proteins, too distantly related to known sequences in databases to retain significant similarity. All proteins, however, probably share common ancestries if one moves far enough back in evolution; therefore, given the huge accumulation of protein sequences in current databases, it could be expected that some proteins with no obvious sequence resemblance to any other share some **residues** that could represent footprints of ancient common ancestries. To identify such putative footprints, the authors searched for short stretches of amino acids present in a given protein sequence that are also found in a significant no. of nonrelated **proteins** in the **database** . The significantly high frequency of occurrence of these patterns in the database would support a common evolutionary source, and a diversity of nonrelated proteins that contain the pattern would express their ancient origin. Using this strategy, significant patterns were found in actual exons exons, but not in randomized amino acid sequences, nor in translated sequences of noncoding DNA, suggesting that this strategy actually leads to the identification of patterns with a biol. significance. These significance patterns are not randomly positioned along the sequences analyzed, but they tend to accumulate within specific regions, producing a profile of discrete domains. In some well-known proteins analyzed in this study, some of these domains are coincident with known motifs. Thus, the procedure described in this paper could be useful for identifying ancient patterns and domains in protein sequences, some of which could also have a **functional** or structural significance.

L1 ANSWER 87 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1995:658821 CAPLUS
DN 123:55029
TI Comparison of methods for searching **protein** sequence **databases**
AU Pearson, William R.
CS Department Biochemistry, University Virginia, Charlottesville, VA, 22908, USA
SO Protein Science (1995), 4(6), 1145-60
CODEN: PRCIEI; ISSN: 0961-8368
PB Cambridge University Press
DT Journal
LA English
AB We have compared commonly used sequence comparison algorithms, scoring matrixes, and gap penalties using a method that identifies statistically significant differences in performance. Search sensitivity with either the Smith-Waterman algorithm or FASTA is significantly improved by using modern scoring matrixes, such as BLOSUM45-55, and optimized gap penalties instead of the conventional PAM250 matrix. More dramatic improvement can be obtained by scaling similarity scores by the logarithm of the length of the library sequence (ln()-scaling). With the best modern scoring matrix (BLOSUM55 or JO93) and optimal gap penalties (-12 for the first **residue** in the gap and -2 for addnl. **residues**), Smith-Waterman and FASTA performed significantly better than BLASTP. With ln()-scaling and optimal scoring matrixes (BLOSUM45 or Gonnet92) and gap penalties (-12, -1), the rigorous Smith-Waterman algorithm performs better than either BLASTP and FASTA, although with the Gonnet92 matrix the difference with FASTA was not significant. Ln()-scaling performed better than normalization based on other simple **functions** of library

sequence length. Ln()-scaling also performed better than scores based on normalized variance, but the differences were not statistically significant for the BLOSUM50 and Gonnet92 matrixes. Optimal scoring matrixes and gap penalties are reported for Smith-Waterman and FASTA, using conventional or ln()-scaled similarity scores. Searches with no penalty for gap extension, or no penalty for gap opening, or an infinite penalty for gaps performed significantly worse than the best methods. Differences in performance between FASTA and Smith-Wterman were not significant when partial query sequences were used. However, the best performance with complete query sequences wa obtained with the Smith-Waterman algorithm and ln()-scaling.

L1 ANSWER 101 OF 116 CAPLUS COPYRIGHT 2003 ACS on STN
AN 1993:142890 CAPLUS
DN 118:142890
TI A systematic search for protein signature sequences
AU Sheridan, Robert P.; Venkataraghavan, R.
CS Med. Res. Div., American Cyanamid Corp., Pearl River, NY, 10965, USA
SO Proteins: Structure, Function, and Genetics (1992), 14(1), 16-28
CODEN: PSFGEY; ISSN: 0887-3585
DT Journal
LA English
AB Signature sequences are contiguous patterns of amino acids 10-50 **residues** long that are assocd. with a particular structure or **function** in proteins. These may be of three types (by the authors nomenclature): superfamily signatures, remnant homologies, and motifs. A systematic search through a database of protein sequences was performed to automatically and preferentially find remnant homologies and motifs. This was accomplished in three steps:. (1); A nonredundant sequence database was generated. (2); BLAST3 (Altschul, S.F.; Lipman, D.J., 1990) was used to generate local pairwise and triplet sequence alignments for every **protein** in the **database** vs. every other. (3); Interesting alignments were selected and grouped into clusters. Most of the clusters contained segments from proteins which share a common structure or **function**. Many of them correspond to signature previously noted in the literature. Three previously recognized motifs are discussed in detail (FAD/NAD-binding, ATP/GTP-binding, and cytochrome b5-like domains) to demonstrate how the alignments generated by the procedure are consistent with previous work and make structural and **functional** sense. Two signatures (for N-acetyltransferases and glycerol-phosphate binding) which have not been previously recognized are also discussed.

```
=> d his

     (FILE 'HOME' ENTERED AT 14:38:25 ON 03 DEC 2003)

     FILE 'CAPLUS' ENTERED AT 14:38:59 ON 03 DEC 2003
L1             9 S (PEPTIDE OR PROTEIN) AND (MOTIF OR CLUSTER) AND SEARCH AND SC
L2            20 S (PEPTIDE OR PROTEIN) AND (MOTIF OR CLUSTER) AND SEARCH AND SC
L3             0 S (PEPTIDE OR PROTEIN) AND ALLIGN? AND (MOTIF OR CLUSTER) AND S
L4             0 S (PEPTIDE OR PROTEIN) AND ALLIGN? AND (MOTIF OR CLUSTER) AND S
L5             0 S (PEPTIDE OR PROTEIN) AND ALIGN? AND (MOTIF OR CLUSTER) AND SE
L6             9 S (PEPTIDE OR PROTEIN) AND ALIGN? AND (MOTIF OR CLUSTER) AND SE
```

Connecting via Winsock to STN

Welcome to STN International! Enter x:x

LOGINID:ssspta1811mxb

PASSWORD:
* * * * * * RECONNECTED TO STN INTERNATIONAL * * * * * *
SESSION RESUMED IN FILE 'USPATFULL' AT 12:15:10 ON 03 DEC 2003
FILE 'USPATFULL' ENTERED AT 12:15:10 ON 03 DEC 2003
CA INDEXING COPYRIGHT (C) 2003 AMERICAN CHEMICAL SOCIETY (ACS)

COST IN U.S. DOLLARS	SINCE FILE ENTRY	TOTAL SESSION
FULL ESTIMATED COST	36.25	143.25

DISCOUNT AMOUNTS (FOR QUALIFYING ACCOUNTS)	SINCE FILE ENTRY	TOTAL SESSION
CA SUBSCRIBER PRICE	0.00	-16.93

=> d his

(FILE 'HOME' ENTERED AT 11:43:00 ON 03 DEC 2003)

FILE 'CAPLUS' ENTERED AT 11:47:26 ON 03 DEC 2003
L1 116 S ((PEPTIDE OR PROTEIN) (2W) DATABASE) AND FUNCTION## AND RESID

FILE 'USPATFULL' ENTERED AT 12:01:09 ON 03 DEC 2003
L2 402 S (DATABASE (3A) PROTEIN (3A) SEQUENCE) (5A) (FUNCTION)
L3 0 S ((DATABASE (3A) PROTEIN (3A) SEQUENCE) (5A) (FUNCTION))/AB,TI
L4 16 S ((DATABASE (3A) PROTEIN (3A) SEQUENCE) AND (FUNCTION))/AB,TI,

=> d 2,9,13,14 bib,abs,kwic

L4 ANSWER 2 OF 16 USPATFULL on STN
AN 2003:266568 USPATFULL
TI Database
IN Swindells, Mark, Easton-on- the- Hill, UNITED KINGDOM
Thornton, Janet, Herts, UNITED KINGDOM
Jones, David, London, UNITED KINGDOM
PI US 2003187587 A1 20031002
AI US 2003-221831 A1 20030204 (10)
WO 2001-GB1105 20010314
PRAI GB 2000-6153 20000314
DT Utility
FS APPLICATION
LREP DARBY & DARBY P.C., P. O. BOX 5257, NEW YORK, NY, 10150-5257
CLMN Number of Claims: 57
ECL Exemplary Claim: 1
DRWN 16 Drawing Page(s)
LN.CNT 3748
AB The invention concerns methods and systems for predicting the **function** of proteins. In particular, the invention relates to databases in which details of sequence homologies, biological **functions** and structures that are shared between proteins of differing sequence have been compiled. The invention also relates to methods, systems and computer software that allows the prediction of protein **function** and structure and, optionally, the ligand binding properties of the proteins within such a database.

AB The invention concerns methods and systems for predicting the **function** of proteins. In particular, the invention relates to databases in which details of sequence homologies, biological **functions** and structures that are shared between proteins of differing sequence have been compiled. The invention also relates to methods, systems and computer software that allows the prediction of protein **function** and structure and, optionally, the ligand binding properties of the proteins within such a database.

CLM What is claimed is:

. . . a combined database; b) comparing each query sequence in the combined database with the other sequences represented in the combined **database** to identify homologous **proteins** or nucleic acid **sequences**; c) compiling the results of the comparisons generated in step b) into a database; and d) annotating the sequences in. . .

. . . separate sequence data resources and one or more structural data resources into a combined database; b) comparing each query protein **sequence** in the combined **database** with the other **protein sequences** represented in the combined database to identify homologous proteins using, for each query sequence: i) one or more pairwise sequence. . .

47) A database system comprising: a **database** of **protein** or nucleic acid **sequence** entries containing sequence information, optionally structure information, functional annotation, and information relating to the alignment of each sequence in the. . .

51) A computer apparatus for predicting the biological **function** of a protein comprising: a processor means comprising: a computer memory for storing a specific sequence of amino acid residues;. . . command a list of proteins with which said specific sequence of amino acid residues is predicted to share a biological **function**.

53) A computer-based system for predicting the biological **function** of a protein comprising the steps of: a) inputting a query sequence of amino acids whose **function** is to be predicted into a database according to either claim 46 or claim 47, or generated according to a. . . to said query sequence, and c) outputting said related sequences in order of similarity with the query sequence, wherein the **functions** of the related sequences correspond to the **functions** predicted for the query sequence.

54) A computer-based system for predicting the biological **function** of a protein comprising the steps of: a) accessing a database according to claim 46 or claim 47, b) inputting a query sequence of amino acids whose **function** is to be predicted into said database; c) interrogating said database for sequences that are similar to said query sequence, and d) outputting said related sequences in order of similarity with the query sequence, wherein the **functions** of the related sequences correspond to the **functions** predicted for the query sequence.

55) A computer system for predicting the biological **function** of a protein, comprising: a central processing unit; an input device for inputting requests; an output device; a memory; at. . . output device; the memory storing a module that is configured so that upon receiving a request to predict the biological **function** of a protein, it performs the steps listed in any one of claims 1-45.

56) A computer-based method for predicting the biological **function** of a protein, comprising the steps of: a) accessing the database of claim 46 or 47, at a remote site, b) inputting into said database a query sequence of amino acids whose **function** is to be predicted; c) interrogating said database for sequences that are similar to said query sequence, and d) presenting said related

sequences in order of similarity with the query sequence, wherein the **functions** of the related sequences correspond to the **functions** predicted for the query sequence.

. . . the computer program mechanism comprising a module that is configured so that upon receiving a request to predict the biological **function** of a protein, it performs a method as recited in any one of claims 1-45.

L4 ANSWER 9 OF 16 USPATFULL on STN
AN 2002:302211 USPATFULL
TI Domain specific knowledge-based metasearch system and methods of using
IN Kincaid, Robert, Half Moon Bay, CA, UNITED STATES
Handley, Simon, Palo Alto, CA, UNITED STATES
Vailaya, Aditya, Los Altos, CA, UNITED STATES
Chundi, Parvathi, Cupertino, CA, UNITED STATES
PI US 2002169764 A1 20021114
AI US 2001-33823 A1 20011219 (10)
PRAI US 2001-289927P 20010509 (60)
DT Utility
FS APPLICATION
LREP Agilent Technologies, Inc, Legal Department, DL429, Intellectual Property Administration, P.O. Box 7599, Loveland, CO, 80537-0599
CLMN Number of Claims: 54
ECL Exemplary Claim: 1
DRWN 8 Drawing Page(s)
LN.CNT 1254
AB A system and method for performing domain-specific knowledge based metasearches. A metasearch engine is provided for accessing a searching text-based documents using generic search engines while simultaneously being able to access publication based databases and sequence databases as well as in-house proprietary databases and any database capable of being interfaced with a web interface so as to produce search results in text format. A data mining module is also provided for organizing raw data obtained by unsupervised clustering, simple relevance ranking, and categorization, all of which are done independently of one another. The system is capable of storing previous search data for use in query refinement or subsequent searches based upon the stored data. A search results collection browser may be provided for analyzing current browsing patterns of the user for developing weighting factors to be used in ordering the results of future searches.

CLM What is claimed is:
. . . of claim 14, wherein each word automatically selected for the generation of the word lists is identified based on a **function** computed from a frequency of occurrence of the word in the particular category for which it is selected, relative to. . .
. . . therefrom, the method comprising the steps of: providing a metasearch engine capable of accessing generic, web-based search engines, publication sites, **sequences** sites, **protein** structure **databases** and pathway information databases; receiving a query inputted by a user to the metasearch engine and searching for documents on a selected set of the generic, web-based search engines, publications sites, **sequences** sites, **protein** structure **databases** and pathway information databases which are relevant to the query; fetching raw data search results in the form of text. . .
. . . of claim 41, wherein each word automatically selected for the generation of the word lists is identified based on a **function** computed from a frequency of occurrence of the word in the particular category for which it is selected, relative to. . .

L4 ANSWER 13 OF 16 USPATFULL on STN

AN 2002:89729 USPATFULL
TI Application of protein structure predictions
IN Benner, Steven Albert, 1501 NW. 68th Ter., Gainesvillle, FL, United States 32605
PI US 6377893 B1 20020423
AI US 1997-914375 19970819 (8)
RLI Continuation-in-part of Ser. No. US 1992-857224, filed on 25 Mar 1992
DT Utility
FS GRANTED
EXNAM Primary Examiner: Shah, Kamini
CLMN Number of Claims: 23
ECL Exemplary Claim: 1
DRWN 2 Drawing Figure(s); 2 Drawing Page(s)
LN.CNT 2630
CAS INDEXING IS AVAILABLE FOR THIS PATENT.
AB A method for making a model for the folded structure of a set of proteins from an evolutionary analysis of a set of aligned homologous protein sequences was claimed in Ser. No. 07/857,224. The instant application concerns methods for using these models. The first method is used to confirm or deny a hypothesis that two proteins are homologous, and is comprised of comparing a predicted structure model for one family of proteins with a predicted structure model for a second family of proteins, or an experimental structure for the second family, and deducing the presence or absence of homology based on the presence or absence of structural similarity flanking key residue motifs in the polypeptide sequence. The second method identifies mutations during the divergent evolution of a protein sequence that are potentially adaptive by identifying episodes during the divergent evolution of a family of proteins where there is a high absolute rate of amino acid substitution, or a high ratio of non-silent substitutions to non-silent substitutions. Amino acids that are changing during this episode are likely to be adaptive. The third is a method for identifying specific in vitro properties of the protein that are likely to play a physiological role in vivo in an organism. This methods involves synthesizing in the laboratory proteins having the reconstructed amino acid sequences of a protein before and after a period of rapid sequence evolution that characterizes adaptive substitution, measuring the in vitro properties of the protein before the episode of rapid sequence evolution, and then measuring the in vivo properties of the protein after the episode of rapid sequence evolution. The in vitro behaviors that remained unchanged through this episode are not likely to have adaptive significance physiologically. The in vitro behaviors that changed through this episode are likely to have adaptive significance physiologically. The fourth concerns method for organizing genome sized sequence databases.

CAS INDEXING IS AVAILABLE FOR THIS PATENT.
CLM What is claimed is:
16. A process for constructing a **database** of **protein sequences** comprised of (a) identifying families of homologous **protein sequences** within said **database**, (b) constructing for each family a multiple sequence alignment, an evolutionary tree, and ancestral sequences at nodes in the tree,. . .
17. A process for the identification of in vitro behaviors of proteins that contribute to their physiological **function**, comprised of (a) identifying branches in an evolutionary tree describing the evolution of the family of related protein that have. . . (d) measuring in the laboratory the behaviors of the ancestral proteins before, during, and after the evolution of new biological **function**, and (e) determining which behaviors change rapidly during this episode to generate as a useful and practical result a list. . .

L4 ANSWER 14 OF 16 USPATFULL on STN
AN 2002:78766 USPATFULL

TI METHODS FOR IDENTIFYING COMPOUNDS THAT BIND TO HLA MOLECULES AND USE OF SUCH COMPOUNDS AS HLA-AGONISTS OR ANTAGONISTS

IN RICHERT, JOHN R., CHEVY CHASE, MD, UNITED STATES
LIU, MING, FALLS CHURCH, VA, UNITED STATES
WU, XIONG-WU, FALLS CHURCH, VA, UNITED STATES
WANG, SHAOMENG, MCLEAN, VA, UNITED STATES
KOHLER, NIKLAS, BRAUNSCHWEIG, GERMANY, FEDERAL REPUBLIC OF
YIN, DAXU, BALTIMORE, MD, UNITED STATES

PI US 2002042423 A1 20020411
AI US 1999-301339 A1 19990429 (9)
PRAI US 1998-83426P 19980429 (60)
DT Utility
FS APPLICATION
LREP ROBIN L. TESKIN, SHAW PITTMAN, 2300 N STREET, N.W., WASHINGTON, DC, 20037-1128
CLMN Number of Claims: 25
ECL Exemplary Claim: 1
DRWN 15 Drawing Page(s)
LN.CNT 1219

CAS INDEXING IS AVAILABLE FOR THIS PATENT.

AB A novel method for identifying compounds which bind HLA molecules and which can be used as HLA agonists or antagonists is provided. These compounds are useful especially in the treatment of autoimmune diseases, transplantation, graft-vs-host disease, and more particularly multiple sclerosis.

CAS INDEXING IS AVAILABLE FOR THIS PATENT.

CLM What is claimed is:

. . . score and forming a group of compounds having a geometrical-fit rank of a predetermined value or higher; minimizing an energy **function** describing interactions between a compound in the group and the receptor site by adjusting coordinates of said compound to obtain. . .

12. The method of claim 11, wherein said energy **function** comprises a Van der Waals interaction term and an electrostatic interaction term.

13. The method of claim 12, wherein minimizing said energy **function** comprises probing said compound's conformational flexibility.

14. The method of claim 13, wherein said minimum energy is a global minimum of said **function**.

16. The method of claim 15, wherein the host molecule is a protein having a known primary structure defined by. . . of the center of each atom of the host molecule comprises: aligning said sequence of the host molecule with a **sequence** of a homologous **protein** obtained from a **database** of proteins having a known tertiary structure, assigning a sequence-homology score to each homologous protein indicating the percentage of amino-acids. . . in said homologous protein to provide a refined tertiary structure having a low energy value defined by an internal energy **function** describing interactions between the atoms of the host molecule in said refined tertiary structure.

. . . the receptor site, said template compound having known binding properties to the host molecule and adding to said internal energy **function** a term describing interactions between said template compound and a side chain of the host molecule.

18. The method of claim 17, wherein said energy **function** comprises a Van der Waals interaction term and a coulombic interaction term.

19. The method of claim 18, wherein minimizing said energy **function** comprises probing said compound's conformational flexibility.

20. The method of claim 19, wherein said minimum energy is a global minimum of said **function**.